Fluidity in the perception of auditory speech

Burgering, Merel; van Laarhoven, Thijs; Baart, Martijn; Vroomen, Jean

Published in:
The Quarterly Journal of Experimental Psychology

DOI:
10.1177/1747021819900884

Publication date:
2020

Document Version
Peer reviewed version

Title

Fluidity in the perception of auditory speech: Cross-modal recalibration of voice gender and vowel identity by a talking face

Authors

Merel A. Burgering[1], Thijs van Laarhoven[1], Martijn Baart[1,2] & Jean Vroomen[1]

[1] Department of Cognitive Neuropsychology, Tilburg University, Warandelaan 2, P.O. Box 90153, 5000 LE Tilburg, the Netherlands

[2] BCBL. Basque Center on Cognition Brain and Language, Donostia - San Sebastián, Spain

Corresponding author

Jean Vroomen

Email address: j.vroomen@uvt.nl

Work phone number: +31 134662394

Abstract

Humans quickly adapt to variations in the speech signal. Adaptation may surface as *recalibration*, a learning effect driven by error-minimization between a visual face and an ambiguous auditory speech signal, or as *selective adaptation*, a contrastive aftereffect driven by the acoustic clarity of the sound. Here, we examined whether these aftereffects occur for vowel identity and voice gender. Participants were exposed to male, female, or androgynous tokens of speakers pronouncing /e/, /ø/, (embedded in words with a consonant-vowel-consonant structure), or an ambiguous vowel halfway between /e/ and /ø/ dubbed onto the video of a male or female speaker pronouncing /e/ or /ø/. For both voice gender and vowel identity, we found assimilative aftereffects after exposure to auditory ambiguous adapter sounds, and contrastive aftereffects after exposure to auditory clear adapter sounds. This demonstrates that similar principles for adaptation in these dimensions are at play.

Keywords: audiovisual integration, gender, vowel, recalibration, selective adaptation

Introduction

Humans constantly integrate different types of sensory input to form coherent representations of the world. This is particularly relevant in social interactions, in which we quickly combine the voice we hear with the face we see when watching our interlocutor. In less than half a second, audiovisual integration processes are initiated that, for example, support perception of the speaker's biological sex – here referred to as gender – (Latinus, VanRullen, & Taylor, 2010), emotion (de Gelder & Vroomen, 2000), and phonetic detail of the spoken input (Baart, Lindborg, & Andersen, 2017; Klucharev, Möttönen, & Sams, 2003; Pilling, 2009; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007; Stekelenburg & Vroomen, 2007; Sumby & Pollack, 1954; van Wassenhove, Grant, & Poeppel, 2005).

Visual information is helpful to classify voice gender because there is substantial variability in the acoustic parameters that contribute to voice gender (i.e., fundamental frequency, (F0), corresponding to the perceived pitch, (Fenn et al., 2011; Pernet & Belin, 2012; Titze, 1989). Seeing the speaker's face while hearing their voice facilitates categorization of both voice and face gender in terms of response times (Joassin, Maurage, & Campanella, 2011). Also, when facial gender is incongruent with the voice, effects are detrimental rather than facilitatory (Huestegge & Raettig, 2018).The effect of seeing a face on voice gender categorization are also stronger than the effect of hearing a voice on face categorization, suggesting that visual information is more dominant in face-voice gender integration than auditory information (Latinus et al., 2010).

Although audiovisual incongruent stimulus materials can contribute to our understanding of multi-sensory integration, it is not clear whether these effects are

caused by a genuine perceptual change or by a response bias. For example, an incorrect voice gender response – such as identifying a female voice as 'male' when it is presented in combination with a male face – may be caused by visual 'capture' (participants really perceived a male voice), but it is also possible that participants simply based their response on the visual information only.

Under natural circumstances, large incongruencies between a face and voice (such as hearing a male voice and seeing a female face) are rare, but what is much more common is that there is a small discrepancy between what is heard and seen, typically because one of the two signals is unclear, degraded, or ambiguous. This distinction is important, because when the auditory signal is ambiguous rather than fully incongruent with the visual input, listeners may use visual facial cues to perceptually adjust/recalibrate their voice gender categories, as they do for phonetic boundaries (Bertelson, Vroomen, & de Gelder, 2003; Sumby & Pollack, 1954). This perceptual shift in the auditory modality minimizes the error between the two signals and induces a learning effect that can be measured as an aftereffect in audio-only trials.

In the phonetic domain, this effect was first demonstrated by Bertelson et al. (2003) who exposed listeners to a moderate phonetic audiovisual conflict. Participants saw a speaker who pronounced /aba/ (or /ada/) while an ambiguous speech sound halfway between /aba/ and /ada/ – A? for auditory ambiguous – was delivered simultaneously. Immediately after exposure, listeners indicated whether ambiguous audio-only test sounds were either /aba/ or /ada/. Identification of the ambiguous sounds was shifted towards the previously seen lip-read information, so the same test sound was perceived more likely as /aba/ when the previous exposure contained lip-

read /aba/ videos, and more likely as /ada/ when exposure contained lip-read /ada/ videos. The rationale behind this effect was that during exposure, the perceptual system minimizes the inter-sensory discrepancy by shifting the auditory phonetic boundary, which leads to longer-term assimilative auditory aftereffects. Bertelson et al. (2003) termed the effect phonetic recalibration, which has proven to be a robust phenomenon (Baart, de Boer-Schellekens, & Vroomen, 2012; Baart & Vroomen, 2010; Franken et al., 2017; Keetels, Pecoraro, & Vroomen, 2015; Keetels, Stekelenburg, & Vroomen, 2016; Kilian-Hütten, Vroomen, & Formisano, 2011; van Linden & Vroomen, 2007; Vroomen & Baart, 2009, 2012; Vroomen, Keetels, De Gelder, & Bertelson, 2004; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004).

Typically, in the paradigm described above, a control condition is included in which participants are exposed to visual information that is paired with canonical/clear and congruent speech sounds that lead to selective adaptation (Eimas & Corbit, 1973). Selective adaptation differs from recalibration in two important ways. Although the same visual information is presented during exposure, selective adaptation is in the opposite direction of recalibration (a contrastive aftereffect, so after exposure to audiovisual /aba/, listeners show *less* /aba/-responses during the auditory test). This effect is not driven by an inter-sensory conflict, but by the repeated presentation of the unambiguous speech sound itself, and is thus independent of the visual information (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). Contrastive aftereffects may reflect neural fatigue of hypothetical 'linguistic feature detectors' (Eimas & Corbit, 1973), but it has also been proposed that they reflect a criterion shift (see Vroomen & Baart (2012) for an overview) or neural sharpening (Kleinschmidt & Jaeger, 2011).

Audiovisual recalibration is quite ubiquitous, as it has also been found to occur for the perception of space (Wozny & Shams, 2011), time (Bermant & Welch, 1976; Bertelson & Aschersleben, 1998; Fujisaki, Shimojo, Kashino, & Nishida, 2004; Keetels & Vroomen, 2007; Radeau & Bertelson, 1974; Vroomen, Keetels, et al., 2004), and for the perception of emotional affect (Baart & Vroomen, 2018). Audiovisual recalibration thus may be a domain-general learning mechanism through which the perceptual system makes necessary adjustments whenever confronted with relatively mild inter-sensory conflicts. Here, the critical question was whether audiovisual recalibration also occurs for the perception of voice gender, which has never been demonstrated before, and vowel identity.

Previous studies on phonetic recalibration mostly focused on consonants because consonants have sharper category boundaries than vowels, see for example (Kuhl, 1991). However, there is some evidence that recalibration also occurs for vowels (Franken et al., 2017; Keetels, Bonte, & Vroomen, 2018). Given that identification of voice gender is mainly driven by fundamental frequency of the sound (Gelfer & Mikos, 2005), and fundamental frequency is more discernible in vowels than in consonants, we envisaged that vowels would provide an ideal platform to simultaneously assess aftereffects of gender and vowel identity. We therefore used audiovisual recordings of a canonical low-pitched male speaker and a high-pitched female speaker pronouncing the vowels /e/ and /ø/. These vowels were chosen because they are close in F1/F2 acoustic space, and easy to discriminate when lip-reading because the rounding of /ø/ is clearly visible. The vowels were embedded in the context of two Dutch words with a similar frequency of occurrence (‘*beek*’ [*stream*] and ‘*beuk*’ [*beech*]). These stimuli then allowed

us to investigate recalibration and selective adaptation of vowels and voice gender in a within-participant and within-stimulus design.

We expected to obtain contrastive aftereffects (indicative of selective adaptation) of voice gender if the auditory tokens were clearly from a male or female speaker (Schweinberger et al., 2008; Zäske, Perlich, & Schweinberger, 2016). Assimilative aftereffects of voice gender (indicative of recalibration) have never been demonstrated before, but as in the phonetic domain, we expected assimilation of voice gender to occur if an androgynous voice was combined with a male or female face. Finding an assimilative effect of voice gender is of interest because it would speak to the generality of the phenomenon since perception of voice gender is quite different from perception of phonemes. For example, voice gender is a more or less stable property over time in the speech signal, which is quite different from phonetic information that is very short-lived and variable between, but also *within* speakers. Furthermore, while vowel categorization occurs in a dense multidimensional acoustic space (largely depending of first and second formant, F1 and F2) that is fine-tuned by language-specific rules, voice gender categorization is, arguably, less complex (a binary male/female distinction, mainly based on fundamental frequency) that is largely shaped by the anatomical differences between the male and female vocal apparatus.

Methods

Participants

Thirty students (11 males, 26 right-handed, mean age of 20.6 years, SD = 2.1) from Tilburg University participated in return for course credits or 8 euro/hour[1]. All participants reported normal hearing, had (corrected to) normal vision and were naïve to the stimuli and research question. Participants provided written informed consent, and the study was conducted in accordance with the Declaration of Helsinki. The Ethics Review Board of the School of Social and Behavioral Sciences of Tilburg University approved the experimental procedures (EC-2016.48).

Stimulus material

Auditory material. We selected four artefact-free audiovisual recordings of a male and female native Dutch speaker pronouncing *beek* and *beuk*. The original speech sound *beek* was pronounced as /e/ (the close-mid front unrounded vowel in IPA with F1 = 471 Hz and F2 = 2013 Hz for the male speaker and F1 = 498 and F2 = 2261 for female speaker) and the original speech sound *beuk* was pronounced as /ø/ (the close-mid front rounded vowel in IPA with F1 = 455 Hz and F2 = 1539 Hz for the male speaker and F1 = 485 Hz and F2 = 1734 Hz for the female speaker). Tokens were chosen to have matching duration of their vowels (duration of male /beek/ = 702 ms, duration of /e/ = 192 ms; duration of male /beuk/ = 631 ms, duration of /ø/ = 205 ms; duration of female /beek/ = 580 ms, duration of /e/ = 191 ms; duration of female /beuk/ = 539 ms, duration of /ø/ = 210 ms). In order to minimize other accidental acoustic

[1] The sample size was larger than in previous work from our lab (see e.g. Bertelson et al., 2003), and was chosen without conducting a formal power analysis.

differences between tokens that might serve as a cue for gender or vowel discrimination, we deleted the release of the final consonant /k/ from *beek* and *beuk* (the unvoiced portions) and replaced them by an identical release from /k/ taken from a /beek/ or /beuk/ recording spoken by a different male. These sounds then served as anchors for two male-female gender continua (one for *beuk* and the other for *beek*). They were created using Tandem-STRAIGHT with a step-size of 2% between adjacent tokens (Kawahara et al., 2008). Tandem-STRAIGHT decomposes a speech sound into five sound parameters, namely spectrum, frequency, aperiodicity, fundamental frequency, and time. Each parameter can be adjusted independently. For each speech sound, we manually identified time landmarks (corresponding with the transitions in the spectrogram, such as on- and offsets of the phonation) and frequency landmarks (corresponding with the first three formants in the spectrogram). Morphed stimuli were then generated by re-synthetization based on interpolation (linear for time; logarithmic for F0, frequency and amplitude) (Schweinberger, Kawahara, Simpson, Skuk, & Zäske, 2014).

We also created two *beuk-beek* vowel continua, one for the male speaker and the other for the female speaker in the same way as described before. We used tokens from the morphing continuum from 5-95% with a step size of 5% from the endpoints towards 40 and 60% and step size of 2% to have higher sampling between 40-60%. We ran a pilot study on seven participants to determine the male-female boundaries (40.6 ± 3.3 for the word *beek* [Ae_{gender}?] and 40.8 ± 4.1 for the word *beuk* [$Aø_{gender}$?]), and the *beuk-beek* vowel boundaries (55.8 ± 3.2 for the male speaker [A_{vowel}?male] and 57.1 ± 2.1 for the female speaker [A_{vowel}?female]. The sounds closest to the these boundaries

were designated as the ambiguous exposure stimulus and test sound (40 for Ae_{gender}?; 40 for $Aø_{gender}$?; 56 for A_{vowel}?male and 58 for A_{vowel}?female). In order to have variation in the test sounds, we also used stimuli of +8% and -8% (denoted as A?+1 and A?-1). The ambiguous boundary tokens and their ambiguous neighbors were used across all participants.

Visual material. During exposure, participants saw the video of a male or female speaker pronouncing *beek* or *beuk.* Recordings were framed as frontal headshots. The entire face of the speaker was visible against a neutral black background and measured 17° horizontally (ear to ear) and 20° vertically (hairline to chin). The videos were edited in Adobe Premiere. A single exposure phase contained four repetitions of either the male or female speaker saying *beek* or *beuk*. It contained a fade-in and fade-out of two frames at the start and the end of the video resulting in a total duration ~5.48 sec. The audio (clear or ambiguous) was dubbed onto the videos without any noticeable synchronization error.

Procedure

General. The experiment took place in a dimly lit sound-attenuated room. Instructions and the face of the speaker were presented on a 25-in monitor (BenQ Zowie XL 2540, 240 Hz refresh rate) positioned at eye-level, ~70 cm from the participant's head. The sound was presented through headphones (Sennheiser HD-203) with a peak intensity of 60 dB SPL. The participant responded by pressing one of two buttons on a response box placed in front of the monitor. Participants were instructed to pay attention to the videos displayed on the monitor, which was checked by the experimenter via a live-feed from a camera in the testing booth. These

instructions were repeated during the breaks between tasks, and after 24 consecutive exposure-test blocks within each task.

Voice gender identification after audiovisual exposure.

In order to induce voice gender recalibration, participants were exposed to four repetitions (ISI=425 ms) of one of the four audiovisual exposure stimuli containing an *androgynous* voice saying *beek*/*beuk* dubbed onto a male/female face: $Ae_{gender}?Ve_{male}$, $Ae_{gender}?Ve_{female}$, $Aø_{gender}?Vø_{male}$ and $Aø_{gender}?Vø_{female}$. The exposure phase was immediately followed by a test phase wherein three test sounds were randomly presented, namely the ambiguous voice gender stimulus with the same vowel that was delivered during exposure (henceforth, /A_{gender}?/), and the two close speech morphs on the same continuum /A?-1/ and /A?+1/ (Fig. 1A). After each test sound, participants decided whether the test token was 'male' or 'female' in a 2AFC task with two buttons on a response box. The next test sound was played 250 ms after a button press.

In order to induce voice gender selective speech adaptation, the exact same procedure was used as for recalibration except that the audiovisual exposure stimuli now contained the *clear* and *gender congruent* audio: (instead of androgynous): $Ae_{male}Ve_{male}$, $Ae_{female}Ve_{female}$, $Aø_{male}Vø_{male}$, $Aø_{female}Vø_{female}$ (Fig. 1B). There were twelve repetitions for each unique exposure-test mini-block, all delivered in pseudo-random order, so in total there were 48 exposure-test mini-blocks for gender recalibration, and 48 mini-blocks for gender selective adaptation.

Vowel identification after audiovisual exposure.

To induce vowel recalibration, the same procedures were used as for gender recalibration, except that the four exposure stimuli to assess recalibration were ambiguous with respect to vowel identity: A_{vowel}?maleVe_{male}, A_{vowel}?male$V\text{ø}_{male}$, A_{vowel}?female$V\text{ø}_{female}$ and A_{vowel}?femaleVe_{female} (henceforth A_{vowel}?). The test sounds were A_{vowel}? and two neighboring sounds on the *beuk-beek* contina. The exposure stimuli to assess selective adaptation of vowels were, as in voice gender selective adaptation, the gender- and vowel-congruent audiovisual stimuli containing clear audio: $Ae_{male}Ve_{male}$, $Ae_{female}Ve_{female}$, $A\text{ø}_{male}V\text{ø}_{male}$, $A\text{ø}_{female}V\text{ø}_{female}$.

Aftereffects of gender and vowel were assessed sequentially with block order counterbalanced across participants. Preliminary analyses showed that block order did not have significant effects on voice gender recalibration and selective adaptation effects, Fs ≤ 1.453, ps ≥ .245, or on vowel recalibration and selective adaptation, Fs < .111, ps > .065. There was also no significant effect of participant gender on voice gender recalibration and selective adaptation, Fs ≤ .737, ps ≥ .401, or on vowel recalibration and selective adaptation, Fs ≤ 3.358, ps ≥ .082, so block order and gender of the participant were not further analyzed.

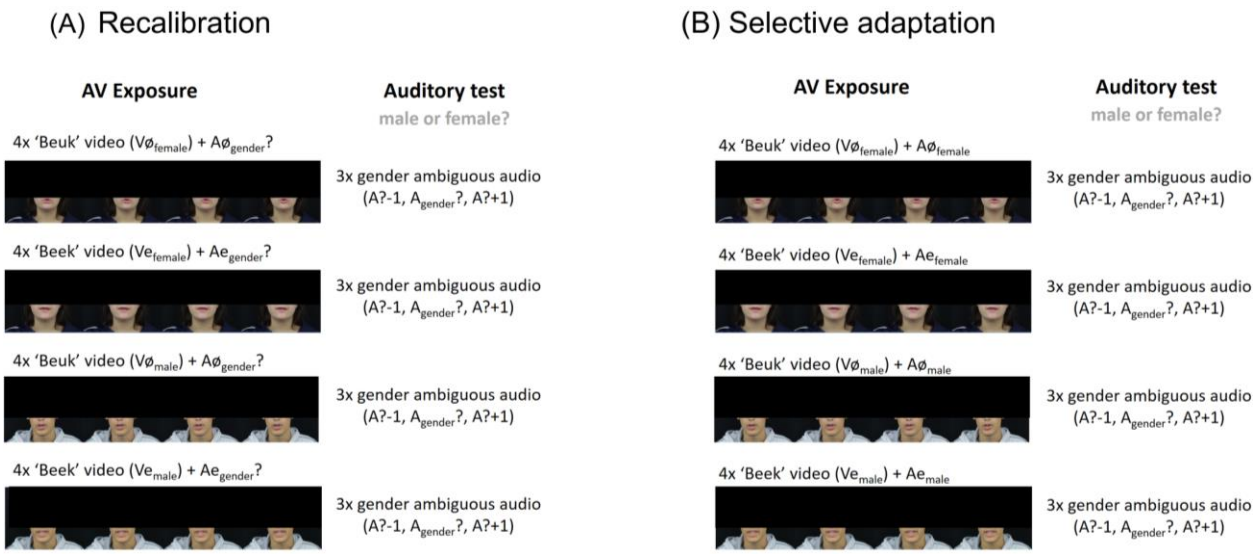

Fig. 1 Overview of the audiovisual exposure-auditory test design. Recalibration (A): four repetitions of a dynamic video of a speaker pronouncing '*beuk*' or '*beek*' combined with audio of ambiguous voice gender were followed by an auditory-only test in which the participant had to categorize the stimulus into the male of female category. Selective adaptation (B): four repetitions of a dynamic video of a speaker pronouncing '*beuk*' or '*beek*' combined with audio of either a male or a female speaker were followed by an auditory-only test in which the participant had to categorize the stimulus into the male of female category. The black bars across the upper half of the faces in the figure were included to anonymize the speakers, but were not presented during the experiment.

Results

Gender recalibration and adaptation

Individual proportions of 'female' responses on the auditory-only test trials were calculated for each combination of Visual exposure gender (female or male), Auditory exposure type (ambiguous or unambiguous), Vowel (/e/ or /ø/), and Test sound (A_{gender}?-1, A_{gender}?, A_{gender}?+1). Data from 9 participants were excluded from the analyses due to unambiguous floor or ceiling effects (see supplementary materials for individual data plots), indicating that they did not adhere to the task instructions or were unable to perform the task correctly. For the remaining 21 participants, grand average proportions of 'female' responses as a function of Visual exposure gender, Vowel, and

Test sound are shown for ambiguous and unambiguous auditory exposure types separately in Figure 2.

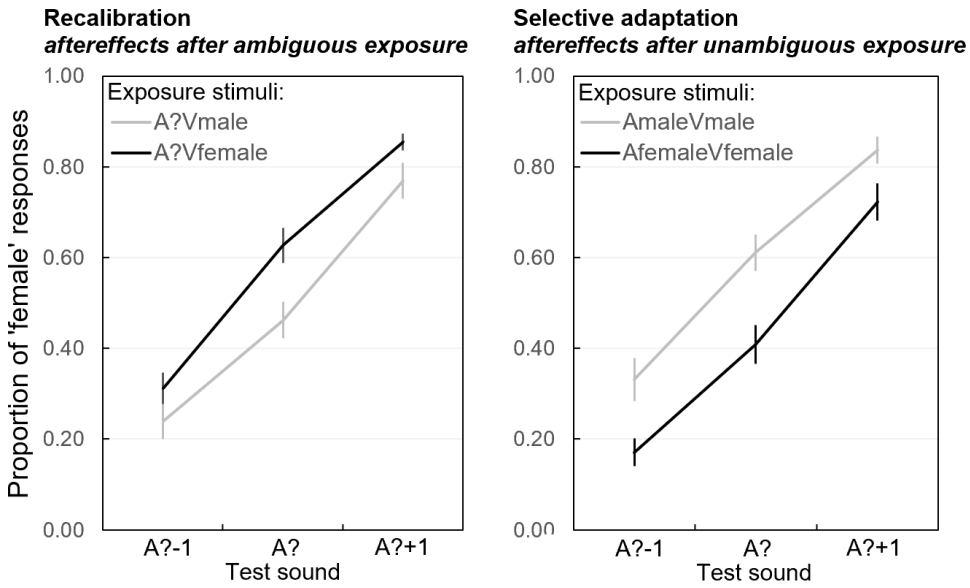

Figure 2. Averaged proportion of 'female' responses on the auditory test that followed AV exposure (N=21) in the Gender identification task, averaged across /e/ and /ø/ vowels. Error bars represent one standard error of the mean.

A generalized linear mixed- effects model with a logistic linking function to account for the dichotomous dependent variable was fitted to the single-trial data (lme4 package in R version 3.5.3). The fitted model included Response (male or female response) as the dependent variable. The model included fixed effects for Visual exposure gender (male or female), Auditory exposure type (ambiguous or unambiguous), Vowel (/e/ or /ø/), and Test sound (A_{gender}?-1, A_{gender}?, A_{gender}?+1), with uncorrelated random intercepts and slopes by participants for the within-participant variables Visual exposure gender and Auditory exposure type, and their interaction. All categorical factors were recoded such that their values were centered around 0. Hence,

the fitted coefficients could be interpreted as the difference in 'female' responses (in log-odds) between two factor levels (e.g. Visual exposure gender male vs female, Auditory exposure type ambiguous vs unambiguous). The fitted model was: Response ~ 1 + VisualExposureGender * AuditoryExposureType * Vowel * TestSound + (1 + VisualExposureGender *AuditoryExposureType | Participant). Fixed effect coefficient estimates are shown in Table 1.

The analysis revealed a main effect of Test sound ($b = 1.36$, $SE = 0.04$, $p < 0.001$), indicative of more 'female' responses to the more female-like test sounds, and a main effect of Auditory exposure type ($b = 0.08$, $SE = 0.03$, $p = 0.01$). Importantly, a significant interaction between Visual exposure gender and Auditory exposure type was found ($b = -0.37$, $SE = 0.09$, $p < 0.001$), indicating that the aftereffects of gender were different for ambiguous and unambiguous auditory exposure stimuli. This interaction effect was further examined with post hoc pairwise contrasts (Bonferroni corrected), testing the effect of visual exposure gender at each auditory exposure type. These contrasts showed a higher proportion of 'female' responses to the test sounds after exposure to ambiguous sounds paired with a visual female speaker, compared to ambiguous sounds paired with a visual male speaker, thereby demonstrating gender recalibration ($b = 0.58$, $SE = 0.18$, $p = 0.001$). In addition, a higher proportion of male responses was reported after exposure to unambiguous sounds paired with a visual female speaker compared to unambiguous sounds paired with a visual male speaker - indicating gender adaptation, $b = -0.91$, $SE = 0.25$, $p < 0.001$).

Table 1. Fixed effect coefficients and standard errors for the fitted mixed effects regression model: Response ~ 1 + VisualExposureGender * AuditoryExposureType * Vowel * TestSound + (1 + VisualExposureGender *AuditoryExposureType | Participant)

Fixed factor	Estimate	Standard error	z-value	*p*
(Intercept)	0.16	0.13	1.242	0.21
VisualExposureGender	0.08	0.06	1.44	0.15
AuditoryExposureType	0.08	0.03	2.56	0.01*
Vowel	-0.02	0.03	-0.66	0.51
TestSound	1.36	0.04	32.74	< 0.001***
VisualExposureGender * AuditoryExposureType	-0.37	0.09	-4.06	< 0.001***
VisualExposureGender * TestSound	-0.03	0.04	-0.76	0.45
VisualExposureGender * Vowel	0.06	0.03	1.78	0.07
AuditoryExposureType * Vowel	0.04	0.03	1.18	0.24
AuditoryExposureType * TestSound	-0.01	0.04	-0.28	0.78
Vowel * Testsound	0.08	0.04	1.99	0.05
VisualExposureGender * AuditoryExposureType * Vowel	-0.04	0.03	-1.21	0.23
VisualExposureGender * AuditoryExposureType * Testsound	0.01	0.04	0.32	0.75
VisualExposureGender * Vowel * Testsound	-0.00	0.04	-0.08	0.94
AuditoryExposureType * Vowel * Testsound	0.01	0.04	0.21	0.83
VisualExposureGender * AuditoryExposureType * Vowel * Testsound	0.05	0.04	1.36	0.17

*$p < .05$; **$p < .01$; ***$p < .001$

Vowel recalibration and adaptation

Individual proportions of /e/ responses on the auditory-only test trials were calculated for each combination of Visual exposure vowel (/e/ or /ø/), Auditory exposure type (ambiguous or unambiguous), Gender (female or male), and Test sound (A_{vowel}?-1, A_{vowel}?, A_{vowel}?+1). Data from 3 participants were excluded from the analyses due to unambiguous floor or ceiling effects (see supplementary materials for individual data plots), indicating that they did not adhere to the task instructions or were unable to perform the task correctly. For the remaining 27 participants, grand average proportions of /e/ responses as a function of Vowel, Visual exposure gender, and Test sound are shown for ambiguous and unambiguous auditory exposure types separately in Figure 3.

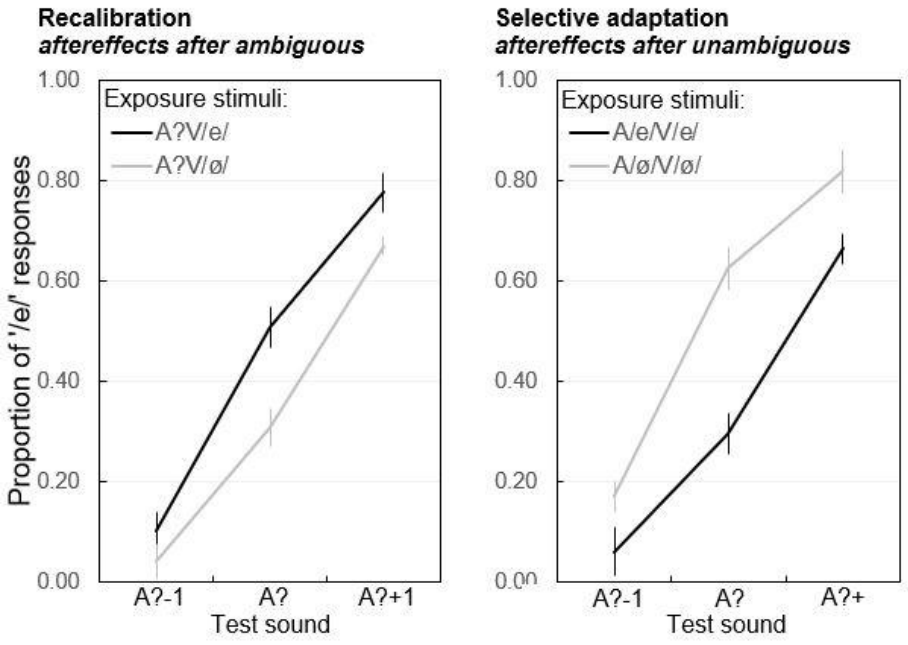

Figure 3. Averaged proportion of '/e/' responses on the auditory test that followed AV exposure (N=27) in the Vowel identification task, averaged across male and female sounds. Error bars represent one standard error of the mean.

A generalized linear mixed- effects model with a logistic linking function to

account for the dichotomous dependent variable was fitted to the single-trial data (lme4 package in R version 3.5.3). The fitted model included Response (/e/ or /ø/ response) as the dependent variable, and fixed effects for Visual exposure vowel (/e/ or /ø/), Auditory exposure type (ambiguous or unambiguous), Gender (female or male), and Test sound (A_{vowel}?-1, A_{vowel}?, A_{vowel}?+1), with uncorrelated random intercepts and slopes by participant for the within-participantvariables Visual exposure vowel and Auditory exposure type, and their interaction. All categorical factors were recoded such that their values were centered around 0. Hence, the fitted coefficients could be interpreted as the difference in /e/ responses (in log-odds) between two factor levels (e.g. Visual exposure vowel /e/ vs /ø/, Auditory exposure type ambiguous vs unambiguous). The fitted model was: Response ~ 1 + VisualExposureVowel * AuditoryExposureType * Gender * TestSound + (1 + VisualExposureVowel *AuditoryExposureType | Participant). Fixed effect coefficient estimates are shown in Table 2.

Table 2. Fixed effect coefficients and standard errors for the fitted mixed effects regression model: Response ~ 1 + VisualExposureVowel * AuditoryExposureType * Gender * TestSound + (1 + VisualExposureVowel *AuditoryExposureType | Participant).

Fixed factor	Estimate	Standard error	z-value	*P*
(Intercept)	-0.52	0.10	-5.38	< 0.001***
VisualExposureVowel	0.11	0.04	2.67	< 0.01**
AuditoryExposureType	-0.12	0.03	-3.62	< 0.001***
Gender	0.25	0.03	8.21	< 0.001***
TestSound	1.79	0.04	42.06	< 0.001***
VisualExposureVowel * AuditoryExposureType	-0.52	0.04	-13.07	< 0.001***

VisualExposureVowel * TestSound	0.00	0.04	0.09	0.93
VisualExposureVowel * Gender	-0.07	0.03	-2.23	0.03*
AuditoryExposureType * Gender	-0.01	0.03	-0.42	0.67
AuditoryExposureType * TestSound	0.03	0.04	0.81	0.42
Gender * Testsound	-0.10	0.04	-2.31	0.02*
VisualExposureVowel * AuditoryExposureType * Gender	0.08	0.03	2.70	< 0.01**
VisualExposureVowel * AuditoryExposureType * Testsound	0.06	0.04	1.49	0.14
VisualExposureVowel * Gender * Testsound	0.04	0.04	0.92	0.36
AuditoryExposureType * Gender * Testsound	-0.02	0.04	-0.60	0.55
VisualExposureVowel * AuditoryExposureType * Gender * Testsound	0.01	0.04	0.36	0.72

*p < .05; **p < .01; ***p < .001

The analysis revealed a negative effect for the intercept (b = −0.52, SE = 0.10, p < 0.001), which indicates a slight overall bias towards /ø/ responses. There was a positive main effect of Test sound (b = 1.79, SE = 0.04, p < 0.001), indicative of more /e/ responses to the more /e/-like test sounds. In addition, there were main effects of Visual exposure vowel (b = 0.11, SE = 0.04, p < 0.01), Auditory exposure type (b = −0.12, SE = 0.03, p < 0.001), and Gender (b = 0.25, SE = 0.03, p < 0.001), and significant interactions between Visual exposure vowel and Gender (b = -0.07, SE = 0.03, p = 0.03), and between Gender and Test sound (b = -0.10, SE = 0.04, p = 0.02). Importantly, a significant interaction between Visual exposure vowel and Auditory exposure type was found (b = -0.52, SE = 0.04, p < 0.001), indicating that the aftereffects of vowel were different for ambiguous and unambiguous Auditory exposure types. Finally, there was a significant interaction between Visual exposure vowel, Auditory exposure type, and Gender (b = 0.08, SE = 0.03, p < 0.01), indicating that the

379 difference in aftereffects of vowel between the ambiguous and unambiguous Auditory
380 exposure types depended on speaker Gender.
381
382 The three-way interaction effect between Visual exposure vowel, Auditory
383 exposure type, and Gender was further examined with post hoc pairwise contrasts
384 (Bonferroni corrected), testing the Visual exposure vowel × Auditory exposure
385 interaction at each level of Gender. These contrasts showed a significant Visual
386 exposure vowel × Auditory exposure interaction for both the male and female speaker
387 (male speaker: b = -1.73, SE = 0.19, p < 0.001, female speaker: b = -2.40, SE = 0.21, p
388 < 0.001). These interaction effects were further explored with post hoc pairwise
389 contrasts (Bonferroni corrected), which showed significant recalibration and adaptation
390 effects for both the male and female speaker. Specifically, a higher proportion of /e/
391 responses to the auditory-only test trials was reported after exposure to ambiguous
392 sounds paired with visual /e/, compared to ambiguous sounds paired with visual /ø/ (i.e.
393 recalibration), male speaker: b = 0.78, SE = 0.13, p < 0.001, female speaker: b = 0.84,
394 SE = 0.14, p < 0.001). In addition, a higher proportion of /e/ responses was reported
395 after exposure to *unambiguous* sounds paired with visual /ø/ compared to unambiguous
396 sounds paired with visual /e/ (i.e. selective adaptation), male speaker: b = -0.96, SE =
397 0.15, p < 0.001, female speaker: b = -1.57, SE = 0.16, p < 0.001).
398 As can be seen in Table 3, vowel recalibration was alike across gender of the
399 exposure stimuli, whereas selective adaptation was larger after female than male
400 exposure stimuli, $t(26) = 2.44$, p = .022.
401

Table 3. Vowel recalibration and selective adaptation per exposure gender, averaged across test-tokens. Aftereffects were quantified as the difference between proportion of /e/-responses after Visual /e/ and Visual /ø/, resulting in *positive* values for recalibration, and *negative* values for selective adaptation. The ambiguous exposure sound A? was ambiguous in terms of vowel identity (not in terms of gender).

Aftereffect type	Exposure gender (Exposure stimulus)	Aftereffect
Recalibration	Male (A?Vmale)	+.12***
	Female (A?Vfemale)	+.12***
Selective adaptation	Male (AmaleVmale)	-.16***
	Female (AfemaleVfemale)	-.24***

*$p < .05$; **$p < .01$; ***$p < .001$ when tested against 0.

Discussion

We found, for the first time, compelling evidence that listeners use the gender of a male or female face to perceptually adjust (recalibrate) their voice gender category boundary, which is presumably based on pitch differences between a male/female voice. When an androgynous voice was dubbed onto the video of a female (instead of male) face during an audiovisual exposure phase, listeners were more likely to categorize an androgynous voice as female in auditory-only posttest trials.

A similar assimilative effect was found for vowels: an ambiguous vowel halfway between /e/ and /ø/ dubbed onto the video of a speaker saying /e/ (instead of /ø/) led to more /e/ responses in auditory-only posttest trials. Gender of the stimulus materials can modulate vowel identification (Johnson, Strand, & D'Imperio, 1999), and we indeed

observed a main effect of Gender on the auditory vowel identification task that followed audiovisual exposure (overall, more /e/ responses were given after exposure to a male rather than female face). Most importantly however, we did not observe a difference in recalibration effect size for vowels induced by male and female exposure materials. We did, however, observe that selective adaptation for vowels was larger after exposure to female adapters rather than male adapters. Johnson et al. (1999) reported that rating female talkers – but not male talkers – as 'stereotypical' is correlated with voice breathiness (in addition to fundamental frequency). Perhaps then, breathiness in the female adapter sound constituted an additional acoustic cue that increased the size of the selective adaptation effect, consistent with the notion that the contrastive adaptation effect is mainly driven by the (unambiguous) exposure sound, and not by the video.

In order to exclude the possibility that assimilative aftereffects were generated by other mechanisms than recalibration (e.g., priming or a simple response strategy to repeat the exposure stimulus), we included a condition in which the exposure stimuli were audio-visually congruent and thus without inter-sensory conflict. With these stimuli, we found in line with previous studies contrastive aftereffects indicative of selective adaptation (Diehl, 1975; Eimas & Corbit, 1973; Schweinberger et al., 2008; Zäske et al., 2016). Selective adaptation of phonetic information is most likely driven by the unambiguous nature of the auditory component of the audiovisual exposure stimulus and appears to be independent of the visual information (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994) The same applies for selective adaptation of voice gender, where the visual information also does not seem to be very relevant. For

example, silent articulating faces did not induce adaptation of perceived auditory gender (Schweinberger et al., 2008).

It remains to be examined in future studies *what* representation listeners adjusted in the case of the gender recalibration task: listeners might have shifted their male/female voice category in general, or only for these two talkers that they heard during the exposure phase. Previous studies on *phonetic* calibration have demonstrated that recalibration is extremely token-specific, and that it even can be ear- and location-specific so that the same ambiguous sound can be simultaneously adapted to two opposing phonetic interpretations if presented in the left and right ear (Keetels et al., 2015). Generalization of recalibration of voice gender, though, might be different. In an informal pilot study (Burgering, Baart, & Vroomen, 2018), we had switched talkers - but not gender - between exposure and test and observed comparable aftereffects. This result, at least tentatively, suggests that voice gender recalibration is not speaker-, or token-specific, but rather generalizes across speakers and tokens.

Another intriguing question for future research is to examine to which extent adaptation in voice gender and voice identity rely on common or separate neural mechanisms. It seems likely that some mechanisms will be shared, while others will be separate. As an example, a study by Green and colleagues (Green, Kuhl, Meltzoff, & Stevens, 1991) provided behavioral evidence that perception of gender and phonetic information rely on dimension-specific mechanisms. The authors showed that the McGurk illusion – such as hearing /da/ when auditory /ba/ is delivered in combination with a face articulating /ga/ – was not modulated by gender incongruency in the audiovisual stimulus, despite the fact that the face-voice gender mismatch was perfectly

clear. Audiovisual integration of phonetic information thus seems to be, at least partially, independent of audiovisual integration of gender information. A reason for this might be that indexical information such as emotional affect or gender is quite holistic in nature and can be acquired from an image or a simple vocalization. In contrast, phonetic processing of speech relies on the fine-grained temporal coherence between what is seen and heard (Cellerino, Borghetti, & Sartucci, 2004; Curby, Johnson, & Tyson, 2012; Lewin & Herlitz, 2002; Sun, Gao, & Han, 2010; Tottenham et al., 2009).

The timing of when gender and phonetic information becomes available, though, might be similar. In an EEG (Electroencephalography) study, Latinus et al. (2010) observed that congruency between facial and vocal gender modulated brain processes within 180 ms and 230 ms after stimulus onset, which aligns with the time-frame during which auditory-only gender differences are processed (Latinus & Taylor, 2012; Zäske, Schweinberger, Kaufmann, & Kawahara, 2009). Interestingly, processing of phonetic congruency is also (partially) realized during this time-window (Arnal, Morillon, Kell, & Giraud, 2009; Baart et al., 2017; Baart, Stekelenburg, & Vroomen, 2014; Stekelenburg & Vroomen, 2007) and audiovisual congruency processing of gender and phonetic information thus overlap in time.

It also remains for future studies to examine whether there is a common neural mechanism for recalibration of voice gender and vowel identity,, especially since there seems to be a good candidate brain region that should be involved in this process: the superior temporal sulcus (STS). Specifically, the STS is involved in lip-read-induced phonetic recalibration (Kilian-Hütten, Valente, Vroomen, & Formisano, 2011), as well as text-induced phonetic recalibration (especially in the right hemisphere, see (Bonte,

Correia, Keetels, Vroomen, & Formisano, 2017), and is also part of a right hemisphere dominated network related to processing vocal gender (Belin et al., 2000; Imaizumi et al., 1997; Von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010), and cross modal integration of face and voice (Blank, Anwander, & von Kriegstein, 2011; Campanella & Belin, 2007; Von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005).

To conclude, humans can flexibly adjust their perceived voice gender categories based on previous exposure. The results are in line with previous studies on voice-face interaction, and the underlying mechanisms seem to operate like those that underlie phonetic selective adaptation and recalibration. The current study inspires future work on the domain general versus domain specific aspects of recalibration.

Acknowledgement

This research was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. The third author was supported by The Netherlands Organization for Scientific Research (NWO: VENI Grant 275-89-027). We thank Alicia Driessen for help with the data collection.

References

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience, 29*(43), 13445-13453. doi:10.1523/JNEUROSCI.3194-09.2009

Baart, M., de Boer-Schellekens, L., & Vroomen, J. (2012). Lipread-induced phonetic recalibration in dyslexia. *Acta Psychologica, 140*(1), 91-95. doi:10.1016/j.actpsy.2012.03.003

Baart, M., Lindborg, A., & Andersen, T. S. (2017). Electrophysiological evidence for differences between fusion and combination illusions in audiovisual speech perception. *Eur J Neurosci, 46*(10), 2578-2583. doi:10.1111/ejn.13734

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53*, 115-121. doi:10.1016/j.neuropsychologia.2013.11.011

Baart, M., & Vroomen, J. (2010). Phonetic recalibration does not depend on working memory. *Experimental brain research, 203*(3), 575-582. doi:10.1007/s00221-010-2264-9

Baart, M., & Vroomen, J. (2018). Recalibration of vocal affects by a dynamic face. *Experimental brain research*, 1-8.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences, 8*(3), 129-135.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309.

Bermant, R. I., & Welch, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills, 43*(2), 487-493. doi:10.2466/pms.1976.43.2.487

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic bulletin & review, 5*(3), 482-489.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*(6), 592-597. doi:10.1046/j.0956-7976.2003.psci_1470.x

Bestelmeyer, P. E., Belin, P., & Grosbras, M. H. (2011). Right temporal TMS impairs voice detection. *Current Biology, 21*(20), R838-R839. doi:10.1016/j.cub.2011.08.046

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice-and face-recognition areas. *Journal of Neuroscience, 31*(96), 12906-12915. doi:10.1523/JNEUROSCI.2091-11.2011

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., & Formisano, E. (2017). Reading-induced shifts of perceptual speech representations in auditory cortex. *Scientific reports, 7*. doi:10.1038/s41598-017-05356-3

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language, 94*, 166-176.

Burgering, M. A., Baart, M., & Vroomen, J. (2018, June 14-17). *Audiovisual recalibration and selective adaptation for vowels and speaker sex.* Paper presented at the 19th International Multisensory Research Forum (IMRF), Toronto, Canada.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in cognitive sciences, 11*(12), 535-543. doi:10.1016/j.tics.2007.10.001

Cellerino, A., Borghetti, D., & Sartucci, F. (2004). Sex differences in face gender recognition in humans. *Brain research bulletin, 63*(6), 443-449. doi:10.1016/j.brainresbull.2004.03.010

Charest, I., Pernet, C., Latinus, M., Crabbe, F., & Belin, P. (2012). Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cerebral Cortex, 23*(4), 958-966.

Curby, K. M., Johnson, K. J., & Tyson, A. (2012). Face to face with emotion: Holistic face processing is modulated by emotional state. *Cognition & Emotion, 26*(1), 93-102. doi:10.1080/02699931.2011.555752

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion, 14*(3), 289-311.

Diehl, R. L. (1975). The effect of selective adaptation on the identification of speech sounds. *Perception & psychophysics, 17*(1), 48-52.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive psychology, 4*(1), 99-109.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2018). The Role of the Human Auditory Corticostriatal Network in Speech Learning. *Cerebral Cortex*.

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *The Quarterly Journal of Experimental Psychology, 64*(7), 1442-1456. doi:10.1080/17470218.2011.570353

Franken, M., Eisner, F., Schoffelen, J., Acheson, D. J., Hagoort, P., & McQueen, J. M. (2017). *Audiovisual recalibration of vowel categories.* Paper presented at the Proceedings of Interspeech 2017.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. Y. (2004). Recalibration on audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773.

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice, 19*(4), 544-554. doi:10.1016/j.jvoice.2004.10.006

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & psychophysics, 50*(6), 524-536.

Huestegge, S. M., & Raettig, T. (2018). Crossing gender borders: bidirectional dynamic interaction between face-based and voice-based gender categorization. *Journal of Voice*.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., . . . Hatano, K. (1997). Vocal identification of speaker and emotion activates differerent brain regions. *Neuroreport, 8*(12), 2809-2812.

Jäncke, L., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *NeuroImage, 15*(4), 733-746.

Joassin, F., Maurage, P., & Campanella, S. (2011). The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study. *NeuroImage, 54*(2), 1654-1661.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of phonetics,, 27*(4), 359-384.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Acoustics, Speech and Signal Processing, ICASSP 2008. IEEE International Conference*, 3933-3936.

Keetels, M., Bonte, M., & Vroomen, J. (2018). A Selective Deficit in Phonetic Recalibration by Text in Developmental Dyslexia. *Frontiers in psychology, 9*.

Keetels, M., Pecoraro, M., & Vroomen, J. (2015). Recalibration of auditory phonemes by lipread speech is ear-specific. *Cognition, 141*, 121-126. doi:10.1016/j.cognition.2015.04.019

Keetels, M., Stekelenburg, J. J., & Vroomen, J. (2016). A spatial gradient in phonetic recalibration by lipread speech. *Journal of phonetics, 56*, 124-130. doi:10.1016/j.wocn.2016.02.005

Keetels, M., & Vroomen, J. (2007). No effect of auditory-visual spatial disparity on temporal recalibration. *Experimental brain research, 182*(4), 559-565.

608 Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the
609 perceptual interpretation of ambiguous sound. *Journal of Neuroscience, 31*(5), 1715-1720.
610 Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2011). Brain activation during audiovisual exposure
611 anticipates future perception of ambiguous speech. *NeuroImage, 57*(4), 1601-1607.
612 doi:10.1016/j.neuroimage.2011.05.043
613 Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and
614 selective adaptation. *Proceedings of the 2nd Workshop on Cognitive Modeling and*
615 *Computational Linguistics*, 10-19.
616 Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-
617 phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain*
618 *Research, 18*(1), 65-75. doi:10.1016/j.cogbrainres.2003.09.004
619 Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the
620 prototypes of speech categories, monkeys do not. *Perception & psychophysics, 50*(2), 93-107.
621 Latinus, M., & Taylor, M. J. (2012). Discriminating male and female voices: differentiating pitch and
622 gender. *Brain topography, 25*(2), 194-204.
623 Latinus, M., VanRullen, R., & Taylor, M. J. (2010). Top-down and bottom-up modulation in processing
624 bimodal face/voice stimuli. *BMC neuroscience, 11*(1), 36.
625 Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition - Women's faces make the difference.
626 *Brain and cognition, 50*(1), 121-128.
627 Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of
628 phonemic perception. *Cerebral Cortex, 15*(10), 1621-1631.
629 Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., & Desai, A. (2013). Neural dynamics
630 of phonological processing in the dorsal auditory stream. *Journal of Neuroscience, 33*(39),
631 15414-15424.
632 Modelska, M., Pourquié, M., & Baart, M. (2019). No “Self” Advantage for Audiovisual Speech
633 Aftereffects. *Frontiers in psychology, 10*(658).
634 Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in*
635 *psychology, 3*, 23.
636 Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of*
637 *Speech, Language, and Hearing Research, 52*(4), 1073-1081.
638 Radeau, M., & Bertelson, P. (1974). The after-effects of ventriloquism. *The Quarterly Journal of*
639 *Experimental Psychology, 26*(1), 63-71.
640 Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is
641 rapidly influenced by context. *Journal of phonetics, 41*(2), 101-116.
642 Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation
643 in speech perception is purely auditory. *Perception & psychophysics, 30*(4), 309-314.
644 Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density
645 electrical mapping and source-analysis of the multisensory mismatch negativity evoked during
646 the McGurk illusion. *Neuropsychologia, 45*(3), 587-597.
647 doi:10.1016/j.neuropsychologia.2006.03.036
648 Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling
649 audiovisual adaptor. *The Journal of the Acoustical Society of America, 95*(6), 3658-3661.
650 Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., . . . Zäske, R.
651 (2008). Auditory Adaptation in Voice Perception. *Current Biology, 18*, 684-688.
652 doi:10.1016/j.cub.2008.04.015
653 Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception.
654 *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 15-25.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically vaid audiovisual events. *Journal of cognitive neuroscience, 19*(12), 1964-1973.

Sugano, Y., Keetels, M., & Vroomen, J. (2016). Auditory dominance in motor-sensory temporal recalibration. *Experimental brain research, 234*(5), 1249-1262. doi:10.1007/s00221-015-4497-0

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.

Sun, Y., Gao, X., & Han, S. (2010). Sex differences in face gender recognition: an event-related potential study. *Brain research, 1327*(69-76).

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85*(4), 1699-1707.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: Judgements from untrained research participants. *Psychiatry Research, 168*(3), 242-249.

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception & Performance, 33*(6), 1483-1494. doi:10.1037/0096-1523.33.6.1483

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181-1186. doi:10.1073/pnas.0408949102

Von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*(1), 48-55.

Von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of cognitive neuroscience, 17*(3), 367-376.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience, 30*(2), 629-638.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110*(2), 254-259. doi:10.1016/j.cognition.2008.10.015

Vroomen, J., & Baart, M. (2012). Phonetic Recalibration in Audiovisual Speech. In M. M. Murray, Wallace, M.T. (Ed.), *The Neural Bases of Multisensory Processes*. Frontiers in Neuroscience: CRC Press/Taylor & Francis.

Vroomen, J., Keetels, M., De Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research, 22*(1), 32-35.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication, 44*, 55-61.

Wozny, D. R., & Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. *Journal of Neuroscience, 31*(12), 4607-4612.

Zäske, R., Perlich, M. C., & Schweinberger, S. R. (2016). To hear or not to hear: Voice processing under visual load. *Attention Perception & Psychophysics, 78*(5), 1488-1495. doi:10.3758/s13414-016-1119-2

Zäske, R., Schweinberger, S. R., Kaufmann, J. M., & Kawahara, H. (2009). In the ear of the beholder: neural correlates of adaptation to voice gender. *European Journal of Neuroscience, 30*, 527-534. doi:10.1111/j.1460-9568.2009/06839.x

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research, 268*, 38-45. doi: 10.1016/j.heares.2010.04.011.

Supplementary materials

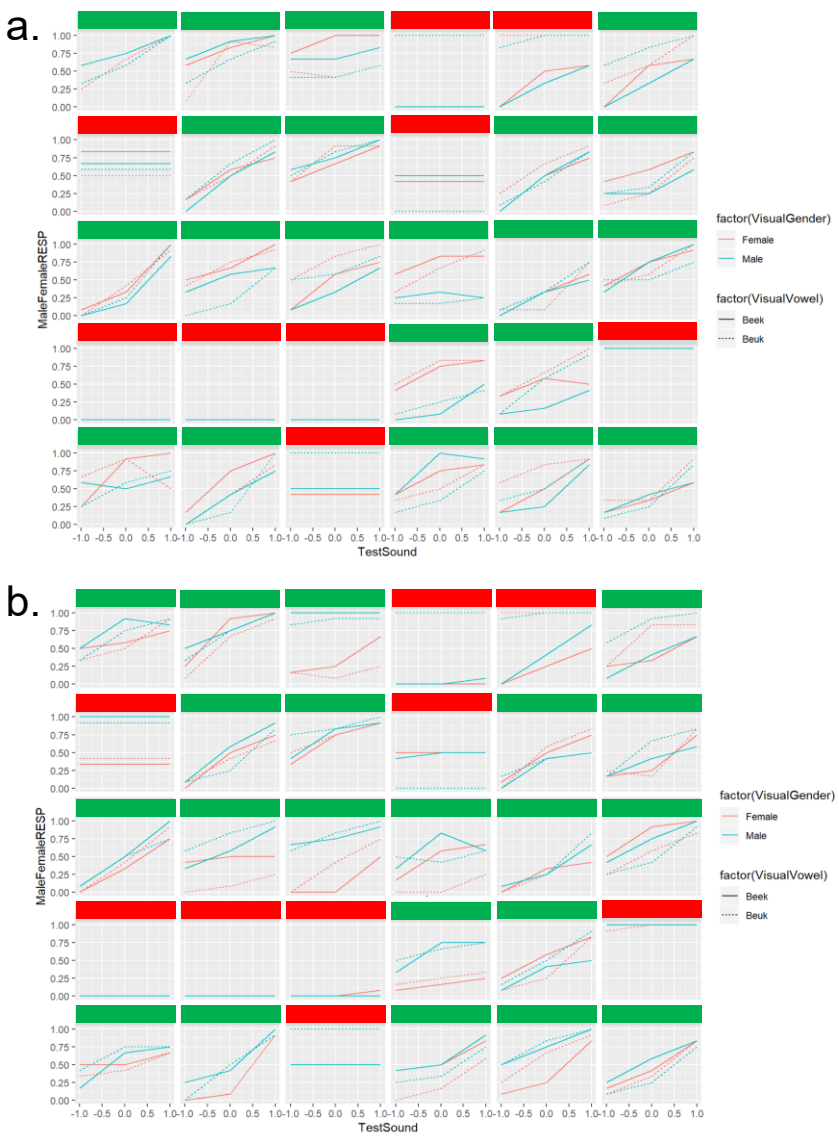

Figure S1. Proportion of female responses in the auditory Gender identification task after AV exposure for all individual participants (N = 30). Participants highlighted by red bars were excluded (N = 9) from the analyses due to ceiling effects (indicating that the test tokens did not represent their perceptual boundaries, and/or participants simply pressed only one key during the test for unknown reasons), or otherwise questionable data patterns. Panel a. represents the data after exposure to ambiguous adapters, panel b. represents the data after exposure to unambiguous adapters.

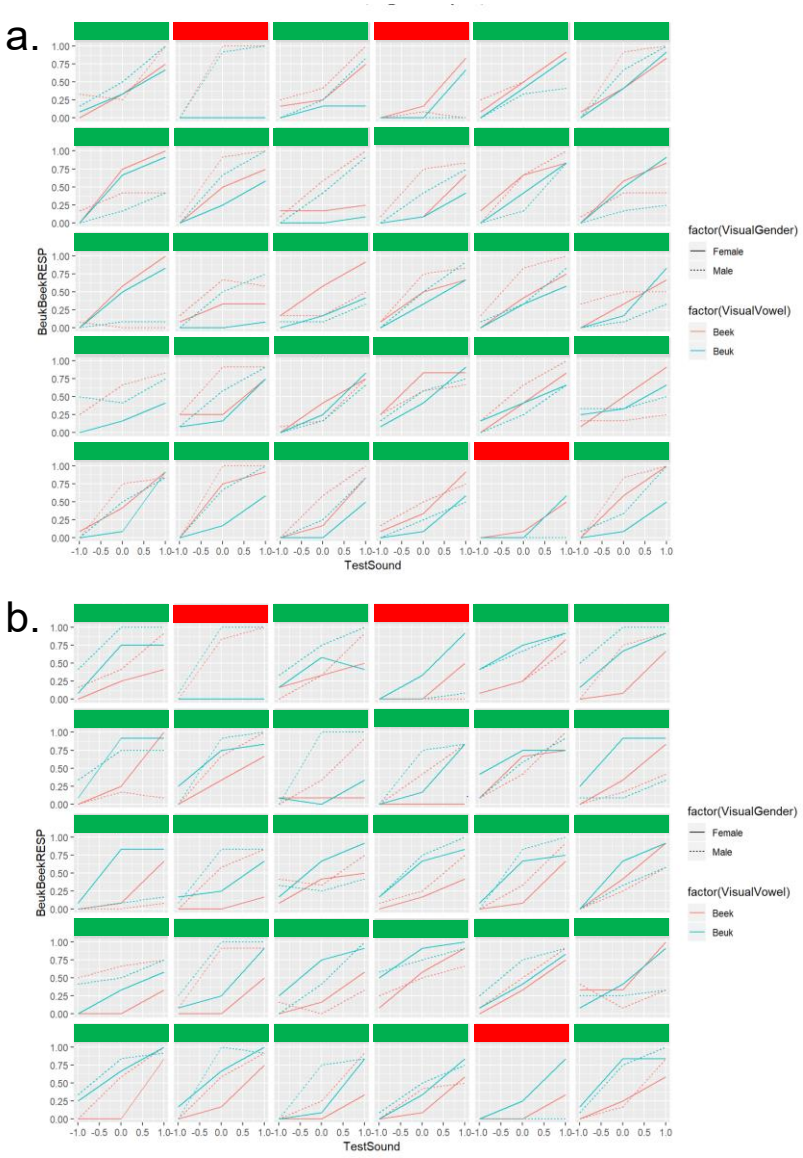

Figure S2. Proportion of /e/ responses in the auditory Gender identification task after AV exposure for all individual participants (N = 30). Participants highlighted by red bars (N = 3) were excluded from the analyses due to ceiling effects (indicating that the test tokens did not represent their perceptual boundaries, and/or participants simply pressed only one key during the test for unknown reasons), or otherwise questionable data patterns. Panel a. represents the data after exposure to ambiguous adapters, panel b. represents the data after exposure to unambiguous adapters.